DOCUMENT RESUME

ED 447 153 TM 031 967

AUTHOR De Ayala, R. J.; Kim, Seock-Ho; Stapleton, Laura M.; Dayton, C. Mitchell

TITLE A Reconceptualization of Differential Item Functioning.

PUB DATE 1999-04-00

NOTE 24p.; Paper presented at the Annual Meeting of the American Educational Research Association (Montreal, Quebec, Canada, April 19-23, 1999).

PUB TYPE Reports - Evaluative (142) -- Speeches/Meeting Papers (150)

EDRS PRICE MF01/PC01 Plus Postage.

DESCRIPTORS *Item Bias; Monte Carlo Methods; *Test Items

IDENTIFIERS Item Bias Detection

ABSTRACT

Differential item functioning (DIF) may be defined as an item that displays different statistical properties for different groups after the groups are matched on an ability measure. For instance, with binary data, DIF exists when there is a difference in the conditional probabilities of a correct response for two manifest groups. This paper suggests that the occurrence of DIF can be explained by recognizing that the observed data do not reflect a homogeneous population of individuals, but are a mixture of data from multiple latent populations or classes. This conceptualization of DIF hypothesizes that when one observes DIF using the current conceptualization of DIF, it is only to the degree that the manifest groups are represented in the latent classes in different proportions. A Monte Carlo study was conducted to compare various approaches to detecting DIF under this formulation of DIF. Results show that as the latent class proportions became more equal, the DIF detection methods identification rates approached null condition levels. (Contains 6 tables, 3 figures, and 27 references.) (Author/SLD)

A Reconceptualization of Differential Item Functioning

R.J. De Ayala
University of Nebraska

Seock-Ho Kim
University of Georgia

Laura M. Stapleton and C.Mitchell Dayton
University of Maryland

Paper presented at the annual meeting of the American Educational Research Association, April 19-23, Montreal, Canada

Abstract

DIF may be defined as an item that displays different statistical properties for different groups after matching the groups on an ability measure. For instance, with binary data DIF exists when there is a difference in the conditional probabilities of a correct response for two manifest groups. This papers argues that the occurrence of DIF can be explained by recognizing that the observed data do not reflect a homogeneous population of individuals but are a mixture of data from multiple latent populations or classes. This conceptualization of DIF hypothesizes that when one observes DIF using the current conceptualization of DIF it is only to the degree that the manifest groups are represented in the latent classes in different proportions. A Monte Carlo study was conducted to compare various approaches to detecting DIF under this formulation of DIF. Results showed that as the latent class proportions became more equal the DIF detection methods identification rates approached null condition levels.

Current approaches to item bias analysis involve the use of differential item functioning (DIF) methods to facilitate the identification of items that are measuring differently for manifest groups. An item identified as exhibiting DIF is subjected to a review by a panel of experts (a.k.a., "logical evidence of bias") to determine whether or not item is biased. It is the panel's conclusion that determines whether an item exhibiting DIF is also biased.

DIF may be defined as an item that displays different statistical properties for different groups after matching the groups on an ability measure (Angoff, 1993). For instance, assuming dichotomous data and an item response theory (IRT) perspective, DIF is defined as a difference in the conditional probabilities of a correct response for two manifest groups (e.g., males and females). Graphically this may be represented as the difference between two item characteristic curves (ICCs) for the same item where one ICC is based on the item parameters estimated from the Reference (R) group response data and the other ICC is based on the parameter estimates based on the Focal (F) group response data after linking the two sets of item parameter estimates.

There are two recognized forms of DIF. In the first case, the ICCs are parallel and this type of DIF is labeled as uniform DIF because one group is favored over the other group across the ability (θ) scale. Nonuniform DIF reflects the fact that members of one manifest group perform better than members of the other manifest group for a part of the θ scale, whereas this relationship is reversed for a different part of the ability scale. Graphically, an item that exhibits nonuniform DIF has ICC for one manifest group crossing the ICC for the other manifest group. Moreover, this allows for the possibility that DIF for one group (positive DIF) may be wholly or partially compensated for by DIF against that group (negative DIF) at another point along the θ continuum.

One interpretation of an item having different ICCs across manifest groups, is that it may be reflective of an item that is measuring ability differently across groups. In effect, the different "abilities" are influencing the probability of a correct response to an item. As such, DIF may be conceptualized as a type of multidimensionality occurring when an item measures multiple dimensions and when the manifest groups differ in their relative locations to one another on the nonprimary ability(ies). If the two groups do not differ in their relative location on the nonprimary dimension(s), then neither group benefits from the nonprimary dimension and DIF does not occur even though the data are multidimensional (cf. Ackerman, 1992).

Current IRT-based DIF analyses create, in essence, two subsamples, albeit with known manifest characteristics (e.g., one subsample is female, the other is male). In this latter case, when the item parameter estimates for an item are not invariant across the manifest groups this difference is interpreted as evidence of DIF. However, in the context of IRT the absence of the invariance of item parameter estimates is evidence of model-data misfit. For instance, if this misfit had occurred across randomly created subsamples (i.e., a large calibration sample is randomly split into two subsamples each of which is separately calibrated), then this would be evidence that the IRT model may not be

appropriate for one or more items. Therefore, DIF is an example of model-data misfit and therefore, regardless of whether the examinee calibration group is randomly split into two subsamples or split into subsamples based on manifest characteristics (e.g., male/female), one may have indication that the model is incorrect for the data at hand. Moreover, items that have been identified as exhibiting DIF (i.e., misfitting items) may still be retained if logical evidence of bias is not forthcoming.

Rather than interpret the difference between, for example, an item's difficulty estimate ($\hat{b}$) for one manifest group and the item's $\hat{b}$ for the other manifest group (after linking) as DIF and trying to explain this DIF with an item bias review, the observed difference may be a reflection of different scales underlying the data. Specifically, the observed data do not reflect a homogeneous population of individuals but are a mixture of data from multiple latent populations or classes (LCs). Within each latent class one has quantitative individual differences (Rost, 1991), but the classes are qualitatively different. Therefore, within a class there is an ability scale and this scale is wholly or in part different than those in other classes. Clearly, there is a multidimensionality in this conceptualization, albeit conceptualized differently than in traditional multidimensional item response theory (e.g., see Ackerman, 1996; Camilli, 1992; Reckase, 1997).

The above conceptualization of the observed data says, in short, that there are one or more latent classes and within each latent class there is an IRT model. In the simplest case there is only one latent class and the calibration sample contains only members from this class and one has model-data fit with a simple IRT model. However, when the observed data consists of members from different latent classes there is not an IRT model that accurately reflects the data for the entire calibration sample (i.e., there is model-data misfit). Rather, there are different item and ability parameters which are conditional on the different subpopulations or latent classes. Mixture distribution models such as those of Rost (1990) and Mislevy and Verhelst (1990) have addressed this general idea and their extensions of the Rasch model have been concerned with solution strategies that differ across subpopulations; also see Kelderman and Macready (1990) for a general framework.

In certain situations examinee samples may actually consist of examinees from different latent classes or subpopulations. In the simplest multiclass situation the examinee sample consists of a mixture of two latent classes. If the latent classes are functionally equivalent to the manifest groups (i.e., 100% of the Reference group are masters, and 100% of the Focal group are nonmasters), then the current conceptualization of DIF will correctly identify DIF. However, it is unlikely that the latent classes will be equivalent to the manifest groups and the two manifest groups may in fact contain members from these two latent classes in different proportions. For example, 80% of the Reference manifest group may consist of masters, whereas 80% of the Focal manifest group may be nonmasters.

Moreover, this conceptualization of DIF hypothesizes that when one observes DIF using the current conceptualization of DIF it is only to the degree that the manifest groups are represented in the latent classes in different proportions. If the manifest groups were equally represented in, say, the two latent

class situation, then it should not be possible to detect DIF using the current IRT-based strategy and these manifest groups. For instance, if the Focal group is defined as black students and 50% come from an impoverished environment (e.g., poor inner city schools) and the balance from a nonimpoverished environment (e.g., affluent suburban schools) the former environment may create a situation in which the students would not have the prerequisite (cognitive) skills to be classified as or belonging to, for example, a masters latent class, whereas the latter condition would lead to the development of the skills that would result in the students being classified in a masters latent class. Similarly, a Reference group consisting of white students could similarly be constructed with the impoverished environment being rural/Appalachia (50%) and the nonimpoverished environment being the suburbs. A comparison of black/white for DIF using standard methods would most likely lead to a false negative conclusion for DIF for each item. However, to the extent that environment affects the development of cognitive skills and thereby affects the classification in a master and nonmaster latent classes, then the items would be performing differentially across classes. That is, the items exhibit DIF, but not with respect to the black/white manifest groups. If one used school/home environment as the characteristic for creating the manifest groups, then DIF would be made evident using standard IRT DIF analysis. Therefore, from this perspective DIF would potentially exist for an item anytime more than one latent class is required to obtain model-data fit for the item set.

The various relationships between two latent classes and manifest groups discussed above are presented in Figure 1:

Insert Figure 1 About Here

In summary, it is believed that the mechanism that gives rise to DIF is best modeled by the assumption of latent classes within which the items can be scaled. The examinee sample consists of a mixture of different latent classes within which one may obtain IRT model-data fit. Given this premise the three scenarios presented above were modeled: the LCs are functionally equivalent to the manifest groups (i.e., 100% of the Reference group belongs to one latent class and 100% of the Focal group belongs to the other latent class), the manifest groups consist of different proportions of the latent classes, and the manifest groups consist of equal proportions of the latent classes. The study consisted of two phases. The first phase compared six different methods for assessing DIF. These methods were the likelihood ratio (G^2) method of Thissen, Steinberg, and Wainer (1988, 1993), the logistic regression method (Swaminathan & Rogers, 1990), Lord's Chi Square (Lord, 1980), Mantel-Haenszel Chi Square as presented by Holland and Thayer (1988), the Exact Signed Area and H Statistic approaches (Raju, 1988, 1990). Thissen et al.'s G^2 approach (G^2_T), Lord's Chi Square (LCS), the Exact Signed Area (ESA) and H statistic are all IRT model-based methods. In the second phase of the study, the logistic regression (LR) and Mantel-Haenszel (MH) methods were used for DIF identification with latent class membership as the classificatory variable for the equal latent class proportions condition.

Methodology

Factors: A fixed test length of 30 items was used and manifest group sizes were 500 simulees for the Focal group (N_F = 500) and 2500 for Reference group (N_R = 2500) (cf. Camilli & Shepard, 1987; Donoghue, Holland, & Thayer, 1993; Zwick, Donoghue, and Grima, 1993). The factors explicitly studied were the number of items exhibiting DIF (NI_{DIF}), the degree to which DIF was expressed by the DIF items (Δb), and the latent class proportions (π_vs). The NI_{DIF} factor consisted of three levels (0%, 10%, and 20% of the 30-item test length) and was included to examine the effect of different degrees of contamination on the conditioning variable on the various DIF approaches. For the DIF items there were two levels of the degree to which DIF was expressed by these items (Δb = 0.3 and Δb = 1.0); these values come from Camilli and Shepard (1987). Two latent class were modeled and the latent class proportions (π_vs) were set at one of three levels: π_1= 0.17/π_2= 0.83,π_1= 0.30/π_2= 0.70, and π_1= 0.50/π_2= 0.50. These three factors resulted in 15 cells and phase 1's design is shown in Figure 2.

Insert Figure 2 About Here

The manifest group sizes crossed by the latent class proportions would theoretically produce the crosstabulations shown in Figure 3.

Insert Figure 3 About Here

Data Generation: Because it was hypothesized that the mechanism underlying DIF can be modeled by an IRT model within latent classes, the data generation required two sets of item parameters, one for each latent class v. The IRT model was the two-parameter logistic (2PL) model with item discrimination fixed at 1.0 (the use a =1.0 has been used previous DIF work, e.g., Camilli & Shepard, 1987). This may be represented as:

$$p(u_i=1 \mid \theta_v, a_{iv}, b_{iv}) = \frac{\exp(a_{iv}(\theta_v - b_{iv}))}{1 + \exp(a_{iv}(\theta_v - b_{iv}))} \qquad (1)$$

where a_{iv}: item i discrimination for latent class v b_{iv}: item i difficulty for latent class v

θ_v: the examinee's latent ability for latent class v u_i = 1: correct response to item i

The b_{iv}s were randomly generated from a N(0,1) distribution. For the DIF conditions the DIF items exhibited DIF as $b_{i1} = b_{i2} + \Delta b$ and for the nonDIF items $b_{i1} = b_{i2}$. Data were generated for each of the 30 items by first assigning simulees to a latent class. This involved summing the πs across the latent classes and then comparing these successive sums to a number randomly generated from a U[0,1] distribution. The first sum that was greater than the random number indicated the simulee's class. Second, a bivariate normal N(0,1, ρ = 0.10) distribution was used to randomly generate a pair of unit

normal deviates to serve as the simulee's θs. The simulees' θs on dimension 1 (v = 1) were rescaled to have a mean of -1.0 and a standard deviation of 1.0 (cf., Zwick, Donoghue, and Grima, 1993), whereas their dimension 2 θs were not rescaled ($\bar{\theta} = 0$, $\sigma_\theta = 1.0$). (Typically in DIF studies, the manifest group members come from populations with different means (e.g., Camilli & Shepard, 1987). However, given the study's premise it the average ability in the classes which needs to vary and manifest group ability means are determined by the mixture of latent class membership.) The simulee's probability of a correct response was calculated according to (1) using the appropriate parameters. If the probability was greater than a number randomly generated from a U[0,1] distribution, then the item response was coded 1 (correct), 0 otherwise. This process was repeated for all 30 items in a data set and for all simulees. Assignment of simulees to Focal and Reference groups was done to match the specifications provided in Figure 1 with the constraints that $N_F = 500$ and $N_R = 2500$. Fifty data sets were generated in this way for each cell.

DIF Assessment: Four IRT-based approaches (G^2_T, LCS, ESA and H Statistic measures) and two nonIRT methods (LR, MH) were used. G^2_T compares a model which assumes a common ICC for both R and F groups with a second model that has all the parameters of the former model plus a group membership parameter. G^2_T assesses whether the group parameter is necessary and one test is conducted for each item in the item set. LR is similar to the G^2_T approach. For LR a set of nested models are created in which the criterion is the item response and the predictors are an ability measure (e.g., the number of correctly answered items, NC), group membership, and the interaction of these two predictors. The simplest of the three models contains only the ability measure, whereas the most complex contains all three predictors; the intermediate model does not contain the interaction term. The three-predictor model is compared to the two-predictor model to see if the interaction term is necessary. Contingent upon the outcome of this test, the two-predictor model is compared to the one-predictor model to determine the necessity of the membership predictor. LCS uses a X^2 to compare the item parameters estimates based on the Reference group data to that obtained using the Focal group data and takes into account the sample variance/covariance of an item's parameter estimates. MH also uses a X^2 test to detect DIF, however, it does not assume the validity of an IRT model. The MH X^2 test is used to determine if responses to an item are independent of group membership after conditioning on a matching variable, such as NC. ESA (signed area) and H (unsigned area) are based on the premise that if the area between the ICC based on the Reference group item parameter estimate(s) and the ICC based on the Focal group item parameter estimate(s) is zero, then the item is functioning "identically" across the two groups. A 'z-test' is used to determine whether any observed nonzero difference for an item is due to randomness or something systematic, such as, group membership. Like LCS, these area measures use the sample variance/covariance matrix of the item parameter estimates.

For performing the G_T^2 the procedure outlined by Thissen, Steinberg, and Wainer (1988, 1993) was followed and MULTILOG (Thissen, 1991) was used for calibration of each of the 50 data sets according to the 2PL model. In a similar fashion, the LR analysis was performed according to Swaminathan and Rogers (1990). A program was written to obtain the MH values for the items. For both MH and LR the number of items correctly answered was used as the ability measure. For LCS, ESA, and H each of the 50 data sets was first decomposed into Reference and Focal groups. Then each group's response data were separately calibrated using the 2PL model; estimates were obtained via BILOG (Mislevy & Bock, 1990). Stocking and Lord's test characteristic curve method (1983), as implemented in EQUATE (Baker, 1993), was used to link the item parameter estimates from the Focal group to that of the Reference group. IRTDIF (Kim & Cohen, 1992) was used to obtain the LCS, ESA, and H values.

Each of the six DIF methods was applied to all 30 items in each data set and the analyses was repeated for each cell in the design. The number of times an item was identified as exhibiting DIF was recorded.

Results

Phase 1:

Table 1 shows the null condition. Across items, the error rates ranged from 3.7% to approximately 8.4% for the various DIF methods. As would be expected, there does not appear to be any differences in the pattern of false positives across levels of π_Vs or across items. LR and MH showed the greatest agreement in their pattern of false positives across all π_Vs conditions ($\pi_1 = 0.17$, $\pi_2 = 0.83$: r = 0.846; $\pi_1 = 0.30$, $\pi_2 = 0.70$: r=0.942; $\pi_1 = 0.50$, $\pi_2 = 0.50$: r=0.911).

Insert Table 1 About Here

Table 2 shows the results for the 'moderate' DIF condition (Δb =0.3); the first three items are the DIF items. The '$\pi_1 = 0.17$, $\pi_2 = 0.83$' condition models the situation in which all manifest group members belong to only one latent class. The LR and MH approaches correctly identify the DIF items 60.7% and 56% of the time, respectively, with false positive rates of approximately 5% or less. As the mixture of latent class membership within manifest group becomes progressively more equal, the LR and MH approaches correct identification rates decreased to null condition levels. LCS, H, ESA, and G_T^2 all had identification rates of 49% or less, although their false positive rates were similar to LR and MH.

Insert Table 2 About Here

The 'high' DIF condition for the three-item level (Table 3) showed high correct identification rates in the '$\pi_1 = 0.17$, $\pi_2 = 0.83$' level for all methods. G_T^2, MH, and LR correctly identified the DIF

items in all 50 replications, although the false positive rates were larger than they were with the Δb =0.3 level. Unlike the pattern in the moderate DIF (NI_{DIF}= 3) condition, the progression from 'π_1 = 0.17, π_2 = 0.83' to 'π_1 = 0.30, π_2 = 0.70' had only a moderate effect on the correct identification rates of the methods, although both ESA and H were more affected than were LCS, G^2_T, MH, and LR. However, for the 'π_1 = 0.50, π_2 = 0.50' condition all of the methods had null condition identification rates. While all methods showed high intercorrelations (i.e., agreements in their patterns of correct identifications and false positives) in the 'π_1 = 0.17, π_2 = 0.83' condition, in the 'π_1 = 0.50, π_2 = 0.50' only LR and MH were highly intercorrelated (r=0.911), with the next highest correlation (r=0.866) existing between MH and G^2_T; LCS and G^2_T had an r= 0.757.

Insert Table 3 About Here

The effect of increased contamination of the matching variable due to the number of DIF items is shown in Tables 4 and 5. While the overall pattern of decreasing correct identification rates as π_1 approached π_2 found in Table 2 exists in Table 4, a comparison of the corresponding correct identification rates shows that they are correspondingly less for the 'π_1 = 0.17, π_2 = 0.83' and 'π_1 = 0.30, π_2 = 0.70' levels in the NI_{DIF}= 6 than in the NI_{DIF}= 3 level. As was the case in Table 3, for the 'π_1 = 0.50, π_2 = 0.50' condition only LR and MH were highly intercorrelated (r=0.899) and the next highest correlation existing between LCS and G^2_T (r=0.878).

Insert Table 4 About Here

Similar to Table 3, Table 5 shows the same overall pattern for all levels of π_Vs, although the false positive rates are larger in the NI_{DIF}= 6 condition than in the NI_{DIF}= 3 condition. The correct identification rates of ESA and H appeared to be more adversely affected by the increase in the number of items exhibiting DIF than were the other measures. Moreover, for the 'π_1 = 0.50, π_2 = 0.50' condition correlation between LR and MH (r=0.921) was largest with the intercorrelation between LCS and G^2 (r=0.746) next.

Insert Table 5 About Here

Phase 2:

The above analyses used the manifest groups for identification of the DIF items. In those conditions where the manifest groups matched the latent class structure, the six methods were, in general, able to identify the DIF items with correct classification rates that were substantially higher than their false positive rates. Overall, MH and LR had consistently higher correct identification rates than did LCS, ESA, and H; G^2_T had rates that were very close and in one instance, better than MH and LR.

Phase 2 involved using the simulees latent class membership in lieu of manifest group membership with LR and MH. LR and MH were selected because they do not assume that an IRT model predicts the

data, and in general, they performed better than or similar to more computationally intensive methods such as G_T^2. Because the most problematic condition for all methods was '$\pi_1 = 0.50, \pi_2 = 0.50$,' this was the condition used in phase 2. (It is assumed, given phase 1 results that the use of latent classes in lieu of manifest group membership for the '$\pi_1 = 0.17, \pi_2 = 0.83$' and '$\pi_1 = 0.30, \pi_2 = 0.70$' conditions will be no worse than for '$\pi_1 = 0.50, \pi_2 = 0.50$'). Furthermore, given the predictable pattern observed in Tables 2 - 5 only three of the possible 5 cells were analyzed: Δb =0.0, Δb =0.30/NI_{DIF}= 3, and Δb =1.0/NI_{DIF}= 6. The results are presented in Table 6. Comparing the false positive rates for the Δb =0.0 level with MH and LRs' rates in Table 1 shows that the rates are higher in Table 6. In contrast, the false positive rates for Δb = 0.3 were comparable to those in Table 2 for MH and LR, while the correct identification rates were substantially higher. Examination of the Δb =1.0 level showed that MH and LR were able to correctly identify the DIF items 100% of the time, however, their false positive rates were 0.420 and 0.480, respectively. These rates were substantially higher than the false positive rates for all DIF methods regardless of condition. Examination of the average MH D-DIF (i.e.,-2.35*log odds ratio (see Dorans & Holland, 1993, p. 41)) for items 7-30 (the false positives) in the '$\pi_1 = 0.50, \pi_2 = 0.50$' condition showed that the degree of DIF would not necessarily be considered important according to the classification criteria used at ETS.

Insert Table 6 About Here

Discussion

Overall, the MH and LR methods had the highest correct identification rates of the six methods compared and were highly linearly related. The G_T^2 method tended to do better than LCS, ESA, and H, particularly in the 'high' degree of DIF conditions. The MH and LR methods may be preferable to the other methods examined because they do not assume an IRT model. Moreover, the G_T^2 method requires as many calibrations as the number of items plus one, whereas LCS, ESA, and H required linking the item parameter estimates for the two manifest groups. While MH examines items for uniform DIF, LR allows for an examination of both uniform and nonuniform DIF.

The '$\pi_1 = 0.17, \pi_2 = 0.83$' level models, according to the study's premise, the mechanism by which DIF exists and can be identified using manifest groups. As π_1 approached π_2 all methods based on manifest groups had greater difficulty in identifying the DIF items. The pattern of decreasing correct identification rates as π_1 approached π_2 found above was expected.

If one conducts a DIF analysis of an item set with respect to gender or race as the manifest groups and determines that none of the items exhibit DIF, then this conclusion of no DIF is with respect only to the manifest groups of gender and race. Such an analysis does not allow one to conclude that there is no DIF in the item set. For instance, 2918 subjects with demographic information were sampled from the data from the National Education Longitudinal Study of 1988 (NELS: 88) and used for a DIF analysis. The "test" consisted of twenty base year math items that spanned the difficulty ($\hat{b}$) range in the math

item pool. A DIF analysis using both MH and LR was performed with these data. Using only examinees which responded that they were "white," two different manifest variables were created. The first was a Residence variable (urban vs suburban) and the second was a School Type variable (public vs private; the latter category included both religious and nonreligious schools). Using Residence as the manifest grouping variable two items were identified as exhibiting DIF by MH and two items were identified by LR, but only one of which was identified by both LR and MH. According to Zieky (1993) all three items would be classified as type A and may still be used for test construction. The LR and MH analyses using School Type as the grouping variable identified the same three items as exhibiting DIF. Two of these would be classified as type A and one of which was type C (MH D DIF=2.37). In this latter case, the class C item would be considered problematic using ETS standards.

In short, to the extent that an item set does not exhibit DIF with respect to race, gender, ethnicity does not mean that there is no DIF within the item set. Other socially identifiable groups may wish to be considered for inclusion as part of routine DIF analyses. In short, the selection of manifest grouping variables is based on political not psychometric considerations.

The above NELS results should not be interpeted as indicating that the math component of NELS contains DIF items because these DIF analyses were performed on a subset (approximately 25%) of the total item pool as well as a subsample of examinees. Conceivably, the use of this smaller item set may affect the conditioning variable with certain DIF methods and therefore the DIF analysis. An item identified as exhibiting DIF with the 20 item set may not exhibit DIF when used with the entire item pool or a different item set. (See Donoghue, Holland, & Thayer (1993) for the conditions under which MH D DIF is not influenced by the number of items in the conditioning variable score.)

Ignoring the issue of self-identification of manifest group membership, a third consideration is the de facto (perhaps innocuous) assumption that individuals within a manifest group are more similar to one another than they are to members of the other manifest group. For instance, there may be a subgroup of the Focal group members that are disadvantaged by one or more items (i.e., in comparison to Reference group members), but the balance of the Focal group is not disadvantaged. In this case, the subgroup is not at all like the majority of the Focal group, however, the relative sizes of the Focal subgoup to the majority of the Focal group may result in the masking of DIF in these items.

Finally, a fourth issue in the current approach to DIF analysis using manifest groups, is that deleting an item because of DIF may have unintended consequences for manifest groups that were not the focus of the analysis (Dorans & Holland, 1993). For example, removing an item that was flagged for positive gender DIF could lower females' scores and raise males' scores while simultaneously increasing the scores of Hispanic and Asian-American females (Dorans & Holland, 1993). Therefore, reformulating DIF in terms of a mixed latent trait- class structure not only allows one to model the current approach DIF, it also minimizes or eliminates some of the above mentioned issues.

References

Ackerman, T. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement, 29,* 67-91.

Ackerman, T. (1996). Developments in multidimensional item response theory. *Applied Psychological Measurement, 20,* 309-310.

Angoff, W.H. (1993). Perspectives on differential item functioning methodology. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Baker, F.B. (1993). EQUATE 2.0: A computer program for the characteristic curve method of IRT equating. *Applied Psychological Measurement, 17,* 20.

Camilli, G. (1992). A conceptual analysis of differential item functioning in terms of a multidimensional item response model. *Applied Psychological Measurement, 16,* 129-147.

Camilli, G. & Shepard, L.A. (1987). The inadequacy of ANOVA for detecting test bias. *Journal of Educational Statistics, 12,* 87-99.

Cleary, T.A., & Hilton, T.L. (1968). An investigation of item bias. *Educational and Psychological Measurement, 28,* 61-75.

Donoghue, J.R., Holland, P.W. , & Thayer, D.T. (1993). A monte carlo study of the factors that affect the Mantel-Haenszel and standardization measures of differential item functioning. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Dorans, N.J., & Holland, P.W. (1993). DIF detection and description: Mantel-Haenszel and standardization. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Holland, P.W., & Thayer, D.T. (1988). Differential item functioning and the Mantel-Haenszel procedure. In H. Wainer & H.I. Braun (Eds.), *Test Validity,* (pp. 129-145). Hillsdale, NJ: Lawrence Erlbaum.

Kelderman, H., & Macready, G.B. (1990). The use of loglinear models for assessing differential item functioning across manifest and latent examinee groups. *Journal of Educational Measurement, 27,* 307-328.

Kim, S.H., & Cohen, A.S. (1992). IRTDIF: A computer program for IRT differential item functioning analysis. *Applied Psychological Measurement, 16,* 158.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems.* Hillsdale, NJ: Erlbaum

Mislevy, R.J., & Bock, R.D. (1982). *BILOG 3: Item analysis and test scoring: with binary logistic model.* Mooresville, IN: Scientific Software, Inc.

Mislevy, R.J., & Verhelst, N. (1990). Modeling item responses when different subjects employ different solution strategies. *Psychometrika, 55,* 195-216.

Raju, N.S. (1988). The area between two item characteristic curves. *Psychometrika, 53,* 495-502.

Raju, N.S. (1990). Determining the significance of estimated signed and unsigned areas between two item response functions. *Applied Psychological Measurement, 14,* 197-207; Correction, *15,* 352.

Reckase, M.D. (1997). Past and future of multidimensional item response theory. *Applied Psychological Measurement, 21,* 25-36.

Rost, J. (1990). A logistic mixture distribution model for polychotomous item responses. *British Journal of Mathematical and Statistical Psychology, 44,* 75-92.

Rost, J. (1991). Rasch models in latent classes: an integration of two approaches to item analysis. *Applied Psychological Measurement, 14,* 271-282.

Stocking, M.L., & Lord, F.M. (1983). Developing a common metric in item response theory. *Applied Psychological Measurement, 7,* 201-210.

Swaminathan, H., & Rogers, J. (1990). Detecting DIF using logistic regression procedures. *Journal of Educational Measurement, 27,* 361-370.

Thissen, D.J. (1991). *MULTILOG* (Version 6.0). Scientific Software, Inc. Mooresville, IN.

Thissen, D.J., Steinberg, L., & Wainer, H. (1988). Use of item response theory in the study of group differences in trace lines. In H. Wainer & H.I. Braun (Eds.), *Test Validity.* Hillsdale, NJ: Lawrence Erlbaum.

Thissen, D.J., Steinberg, L., & Wainer, H. (1993). Detection of differential item functioning using the parameters of item response models. In P.W. Holland & H. Wainer (Eds.), *Differential item functioning* . Hillsdale, NJ: Lawrence Erlbaum.

Zieky, M. (1993). Practical questions in the use of DIF statistics in test development. In P.W. Holland & H. Wainer (Eds.), *Differential Item Functioning.* Hillsdale, NJ: Lawrence Erlbaum.

Zwick, R., Donoghue, J.R., and Grima, A. (1993). Assessment of differential item functioning for performance tasks. *Journal of Educational Measurement, 30,* 233-251.

Table 1: Number of Times Each Item Was Identified as Exhibiting DIF (Δb = 0: False Positives; N=50) and the intercorrelations of the DIF methods

	Latent Class Proportions																	
	$\pi_1 = 0.17, \pi_2 = 0.83$ DIF Method						$\pi_1 = 0.30, \pi_2 = 0.70$ DIF Method						$\pi_1 = 0.50, \pi_2 = 0.50$ DIF Method					
Item	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH
1	2	3	6	2	2	2	2	0	7	2	1	1	2	3	4	4	4	3
2	3	2	4	5	3	3	3	1	4	4	1	1	0	4	6	1	4	4
3	3	1	1	2	3	3	4	4	7	4	3	3	3	3	4	3	2	2
4	4	3	2	2	4	2	0	1	1	0	3	3	0	1	2	0	2	1
5	3	3	4	3	3	3	3	3	5	4	1	1	7	5	8	5	5	3
6	3	3	4	3	2	2	3	2	2	3	2	2	1	0	2	2	0	0
7	1	1	1	0	0	0	2	1	1	2	4	4	0	1	2	0	2	1
8	0	0	0	0	1	0	5	4	8	6	4	4	1	1	3	1	4	4
9	1	1	2	1	0	0	1	1	2	1	0	0	1	1	3	1	4	3
10	1	1	1	1	3	1	1	0	4	3	2	2	3	1	3	4	3	3
11	1	2	5	0	1	1	4	2	3	7	3	3	0	2	2	0	3	2
12	3	2	2	4	1	2	1	0	3	2	4	4	2	1	2	2	1	1
13	1	2	2	2	1	1	4	2	6	5	2	2	3	3	6	4	6	6
14	1	1	5	3	3	1	3	1	4	4	2	1	1	2	3	3	4	3
15	1	0	2	2	2	2	3	2	3	4	2	1	2	3	3	1	0	1
16	1	6	3	3	5	5	1	0	6	1	3	2	1	1	1	1	2	1
17	3	2	2	2	0	0	1	0	2	1	1	1	2	2	2	3	3	1
18	3	3	4	4	3	3	6	4	7	9	7	7	3	2	3	2	2	1
19	4	1	3	3	4	4	4	2	2	5	4	5	2	1	3	2	1	1
20	1	2	2	2	4	3	2	3	5	3	2	3	2	2	6	2	2	2
21	2	4	6	4	2	2	1	1	2	1	3	3	3	3	4	3	3	3
22	5	2	2	3	5	4	0	1	1	0	2	1	3	3	4	3	5	4
23	2	3	4	3	3	2	2	4	6	2	2	2	1	1	2	1	1	1
24	4	3	6	3	0	0	2	3	2	2	3	3	3	1	2	3	2	2
25	2	3	5	4	5	3	1	2	6	1	2	1	2	1	10	3	2	1
26	4	4	5	3	1	1	4	4	5	5	4	3	2	2	2	3	1	1
27	1	1	3	0	1	0	3	5	8	7	6	5	0	0	2	2	0	0
28	2	1	3	2	2	1	2	1	6	2	1	1	3	1	3	3	2	2
29	1	1	4	3	1	2	1	0	2	1	5	5	2	3	4	3	6	5
30	4	3	7	4	2	2	1	2	6	4	5	4	2	3	4	4	3	3

Proportions of False Positives	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.045	0.043	0.067	0.049	0.045	0.037
$\pi_1 = 0.30, \pi_2 = 0.70$	0.047	0.037	0.084	0.063	0.056	0.052
$\pi_1 = 0.50, \pi_2 = 0.50$	0.038	0.038	0.070	0.046	0.053	0.043

Table 2: Number of Times Each Item Was Identified as Exhibiting DIF (Δb = 0.3; N=50)

	Latent Class Proportions																	
	$\pi_1 = 0.17, \pi_2 = 0.83$						$\pi_1 = 0.30, \pi_2 = 0.70$						$\pi_1 = 0.50, \pi_2 = 0.50$					
	DIF Method						DIF Method						DIF Method					
Item	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH
1*	17	10	18	21	28	25	4	4	6	8	6	6	2	1	5	2	2	2
2*	32	26	34	33	37	33	12	8	15	14	11	10	0	3	1	2	2	2
3*	12	3	10	19	26	26	3	2	5	7	7	5	1	0	2	1	1	1
4	3	5	5	1	3	0	7	2	4	6	4	3	1	1	6	1	2	2
5	1	4	5	1	3	1	2	3	3	2	3	2	2	2	3	2	3	3
6	3	6	6	0	2	1	4	4	6	4	4	3	3	4	7	4	5	3
7	3	0	3	3	2	3	3	2	5	6	4	3	4	3	6	4	3	1
8	3	4	5	3	1	1	3	2	5	3	2	3	1	0	2	1	2	2
9	1	3	3	3	1	1	1	1	1	2	1	2	1	1	1	3	3	3
10	2	4	6	4	5	7	5	1	5	7	4	3	2	0	3	2	5	4
11	2	2	4	4	2	1	0	3	3	1	3	3	3	5	6	4	6	5
12	4	2	2	3	4	2	2	2	2	2	3	2	2	1	3	1	1	1
13	1	2	5	1	2	2	2	3	3	2	2	2	1	3	4	1	4	3
14	2	2	2	2	3	3	3	5	6	4	4	4	4	3	6	4	6	5
15	2	0	2	2	2	4	3	1	1	3	3	3	1	1	2	3	2	2
16	2	2	5	4	2	4	4	4	7	6	4	3	7	4	6	6	5	5
17	4	5	8	5	3	3	1	3	5	2	5	5	2	4	6	3	6	5
18	5	5	8	3	4	4	2	6	7	1	2	1	4	1	3	3	2	1
19	1	1	2	2	2	2	2	1	1	2	1	1	3	0	1	4	3	3
20	1	2	4	2	2	2	0	0	2	1	0	0	4	2	7	5	3	3
21	4	4	4	3	4	2	1	2	2	2	1	1	0	3	4	2	4	2
22	6	3	3	3	0	0	10	5	5	7	4	2	1	2	2	0	1	1
23	2	1	2	5	3	3	2	2	2	4	4	3	0	0	0	1	0	0
24	3	4	3	2	4	2	3	2	7	3	2	2	1	1	4	1	0	0
25	1	3	6	2	3	3	1	2	2	1	3	3	2	1	3	3	1	1
26	3	3	4	2	2	1	4	3	7	4	5	5	1	0	3	2	3	3
27	3	4	4	3	4	4	1	3	4	2	2	1	0	2	2	2	2	2
28	2	3	7	3	3	1	3	4	6	3	5	5	1	3	3	2	4	5
29	3	3	5	4	3	3	1	3	5	4	4	3	1	0	1	1	1	1
30	1	4	2	0	3	1	1	2	3	1	2	1	4	3	7	5	4	3

*denotes item with DIF

Proportions Correct	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.407	0.260	0.413	0.487	0.607	0.560
$\pi_1 = 0.30, \pi_2 = 0.70$	0.127	0.093	0.173	0.193	0.160	0.140
$\pi_1 = 0.50, \pi_2 = 0.50$	0.020	0.027	0.053	0.033	0.033	0.033

Proportions of False Positives	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.050	0.060	0.085	0.052	0.053	0.045
$\pi_1 = 0.30, \pi_2 = 0.70$	0.053	0.053	0.081	0.063	0.060	0.051
$\pi_1 = 0.50, \pi_2 = 0.50$	0.041	0.037	0.075	0.052	0.060	0.051

Table 3: Number of Times Each Item Was Identified as Exhibiting DIF (Δb = 1.0; N=50)

	Latent Class Proportions																	
	$\pi_1 = 0.17, \pi_2 = 0.83$						$\pi_1 = 0.30, \pi_2 = 0.70$						$\pi_1 = 0.50, \pi_2 = 0.50$					
	DIF Method						DIF Method						DIF Method					
Item	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH
1*	50	50	50	50	50	50	47	40	48	49	49	49	3	2	7	2	3	3
2*	50	50	50	50	50	50	50	49	50	50	50	50	1	2	7	2	3	4
3*	49	29	37	50	50	50	47	10	32	47	46	46	1	2	3	4	3	2
4	12	11	12	11	10	8	8	7	8	8	9	9	6	3	7	6	4	4
5	9	7	7	6	8	5	6	6	6	5	7	5	4	4	5	5	0	0
6	9	9	10	5	6	5	1	3	5	1	1	0	2	2	6	3	3	2
7	2	1	0	2	2	2	2	2	2	4	5	5	1	1	1	2	3	2
8	5	5	8	7	8	8	4	4	7	4	6	6	0	0	3	0	2	1
9	0	0	0	3	0	2	2	0	2	3	3	2	2	1	1	3	2	1
10	2	1	4	2	3	3	3	2	5	4	4	4	2	2	5	3	2	1
11	5	8	12	6	5	5	4	6	9	5	6	5	4	3	7	3	4	2
12	10	5	6	8	9	8	4	3	1	3	6	5	3	3	2	1	3	3
13	4	8	7	3	5	5	6	8	9	7	4	4	4	2	9	5	4	3
14	7	7	12	9	8	8	2	2	4	1	1	2	4	3	8	3	7	4
15	4	2	5	4	3	3	3	2	2	3	2	3	4	3	6	7	4	4
16	2	2	5	3	4	3	1	1	2	1	2	2	2	3	5	5	4	3
17	3	6	5	3	4	4	2	2	2	4	3	3	2	4	6	4	5	4
18	5	5	6	5	5	4	4	5	7	5	6	4	3	4	7	6	5	5
19	2	1	3	4	3	6	2	3	3	5	5	4	1	0	0	1	2	3
20	8	8	10	7	7	8	4	4	7	4	4	4	1	0	4	2	0	0
21	9	3	4	6	2	2	3	3	5	4	7	7	0	1	2	0	1	1
22	7	5	6	3	5	5	3	3	3	1	2	1	0	1	1	2	1	0
23	4	4	7	6	5	6	3	3	5	4	2	2	2	0	3	3	1	1
24	9	7	8	8	7	5	4	7	7	7	7	5	1	1	2	1	4	4
25	7	8	12	5	7	7	3	3	5	3	3	3	4	3	6	5	3	3
26	10	9	8	5	4	2	3	4	5	3	3	2	1	1	4	2	5	3
27	11	11	12	10	10	9	5	5	9	6	4	4	3	2	4	4	3	3
28	7	9	10	7	4	3	4	4	6	5	5	4	0	0	3	1	0	0
29	2	3	2	4	5	4	2	2	4	5	7	5	0	1	1	0	0	0
30	3	3	3	3	6	6	4	7	8	5	7	7	4	3	6	4	4	4

*denotes item with DIF

Proportions Correct	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.993	0.860	0.913	1.000	1.000	1.000
$\pi_1 = 0.30, \pi_2 = 0.70$	0.960	0.660	0.867	0.973	0.967	0.967
$\pi_1 = 0.50, \pi_2 = 0.50$	0.033	0.040	0.113	0.053	0.060	0.060

Proportions of False Positives	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.117	0.110	0.136	0.107	0.107	0.101
$\pi_1 = 0.30, \pi_2 = 0.70$	0.068	0.075	0.102	0.081	0.090	0.079
$\pi_1 = 0.50, \pi_2 = 0.50$	0.044	0.038	0.084	0.060	0.056	0.045

Table 4: Number of Times Each Item Was Identified as Exhibiting DIF (Δb = 0.3; N=50)

	Latent Class Proportions																	
	$\pi_1 = 0.17, \pi_2 = 0.83$						$\pi_1 = 0.30, \pi_2 = 0.70$						$\pi_1 = 0.50, \pi_2 = 0.50$					
	DIF Method						DIF Method						DIF Method					
Item	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH
1*	14	7	16	22	25	22	6	1	12	10	10	9	0	2	3	0	3	2
2*	19	18	24	23	28	28	4	1	5	4	3	4	2	2	3	2	2	2
3*	11	2	8	24	21	21	5	4	8	7	14	13	3	1	2	4	4	4
4*	9	2	5	15	19	16	3	0	4	5	8	8	1	1	2	1	4	3
5*	9	0	2	12	15	15	2	1	4	5	8	7	2	2	7	3	0	0
6*	11	0	8	13	14	16	4	2	4	5	5	5	3	3	3	5	4	4
7	4	4	2	6	3	3	0	0	2	1	2	1	2	3	3	3	0	0
8	1	0	3	2	1	1	0	1	4	1	1	1	1	2	3	2	1	1
9	0	1	2	2	1	1	0	1	0	1	1	2	3	1	3	4	2	2
10	3	1	4	3	2	2	1	5	7	2	3	2	4	1	4	4	2	1
11	5	5	9	6	4	5	3	3	5	3	3	3	2	2	4	2	3	2
12	3	5	6	1	2	1	3	1	1	5	2	1	4	1	2	4	2	1
13	6	8	10	5	5	4	4	3	6	5	3	3	3	1	5	4	4	3
14	0	1	2	1	1	1	2	3	7	2	1	1	1	1	4	0	2	2
15	2	0	1	3	4	4	1	3	4	4	2	1	1	1	2	2	2	1
16	3	5	9	3	4	6	0	3	4	0	4	4	2	1	3	2	1	1
17	3	4	10	5	4	2	4	6	9	4	5	5	2	3	5	3	3	2
18	6	9	13	5	5	5	3	3	6	3	4	3	3	1	5	3	2	2
19	1	1	2	3	3	3	3	0	2	6	2	2	2	3	4	2	2	2
20	4	4	8	6	5	5	1	2	3	2	2	2	2	2	2	3	2	2
21	9	7	8	5	5	5	1	4	5	0	2	2	0	1	2	2	2	2
22	5	4	4	2	2	1	6	3	3	4	3	2	2	0	0	3	4	4
23	3	3	6	2	2	1	1	1	4	0	2	1	1	1	3	1	2	1
24	4	2	3	5	2	2	4	6	10	4	4	4	1	0	5	1	0	0
25	6	6	9	6	4	4	2	2	6	3	3	3	2	3	6	2	3	3
26	4	3	6	4	8	3	2	4	4	2	6	4	5	4	4	4	4	3
27	3	3	4	3	4	3	0	0	2	0	1	2	1	3	1	2	1	1
28	4	5	8	3	4	5	2	1	4	2	1	1	7	3	11	7	2	3
29	3	4	5	4	3	3	1	6	2	4	5	4	3	3	4	4	5	5
30	3	6	5	1	0	0	3	8	8	2	6	4	1	3	3	2	3	4

*denotes item with DIF

Proportions Correct	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.243	0.097	0.210	0.363	0.407	0.393
$\pi_1 = 0.30, \pi_2 = 0.70$	0.080	0.030	0.123	0.120	0.160	0.153
$\pi_1 = 0.50, \pi_2 = 0.50$	0.037	0.037	0.067	0.050	0.057	0.050

Proportions of False Positives	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.071	0.076	0.116	0.072	0.065	0.058
$\pi_1 = 0.30, \pi_2 = 0.70$	0.039	0.058	0.090	0.050	0.057	0.048
$\pi_1 = 0.50, \pi_2 = 0.50$	0.046	0.037	0.073	0.055	0.045	0.040

Table 5: Number of Times Each Item Was Identified as Exhibiting DIF (Δb =1.0; N=50)

	Latent Class Proportions																	
	$\pi_1 = 0.17, \pi_2 = 0.83$						$\pi_1 = 0.30, \pi_2 = 0.70$						$\pi_1 = 0.50, \pi_2 = 0.50$					
	DIF Method						DIF Method						DIF Method					
Item	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH	LCS	ESA	H	G^2	LR	MH
1*	50	48	50	50	50	50	45	29	43	46	49	49	2	5	5	2	2	2
2*	50	50	50	50	50	50	47	48	47	47	48	48	3	2	5	3	5	4
3*	47	9	23	50	50	50	36	9	29	42	45	44	6	2	3	6	3	3
4*	48	8	28	50	50	50	37	8	25	36	41	40	2	2	3	2	0	0
5*	48	5	23	49	50	50	35	3	26	38	45	44	1	3	6	2	1	1
6*	48	11	37	50	50	50	44	10	31	43	48	48	2	3	1	5	7	6
7	3	3	3	4	5	6	5	1	2	5	5	5	4	2	3	2	2	2
8	10	6	19	15	10	13	4	6	7	8	8	7	1	2	5	3	1	1
9	2	0	0	4	4	4	4	0	4	7	6	6	6	3	5	6	4	4
10	6	5	8	8	7	9	4	3	6	10	6	5	4	1	3	5	2	2
11	12	17	19	12	11	12	3	5	8	4	5	5	1	0	5	1	2	0
12	13	10	9	5	15	11	6	4	4	5	4	4	1	1	4	4	3	3
13	16	17	18	17	19	17	4	7	7	5	6	4	3	2	6	5	3	3
14	8	7	9	9	10	9	5	3	8	5	4	3	0	0	2	0	1	1
15	4	1	3	5	4	5	2	2	4	2	3	2	2	0	2	6	6	6
16	9	12	18	11	9	8	4	6	9	4	6	6	4	4	7	5	4	4
17	15	19	20	17	14	15	9	10	13	14	9	9	5	4	6	4	4	4
18	14	15	15	15	15	13	6	6	8	9	8	8	2	3	6	3	4	3
19	2	0	2	7	7	7	2	1	3	3	3	4	3	0	1	3	2	2
20	15	20	18	15	11	9	2	9	12	5	11	8	2	2	3	4	2	2
21	12	11	10	8	12	7	6	3	3	4	6	4	0	0	2	0	2	1
22	17	9	9	5	17	7	2	4	5	0	2	1	4	3	3	4	3	2
23	11	10	15	12	11	10	8	10	13	10	9	9	3	2	5	4	1	1
24	14	15	16	14	11	10	7	7	10	6	10	8	2	0	3	2	2	2
25	15	17	19	16	17	16	7	10	12	9	8	7	1	1	5	1	0	0
26	14	9	11	12	14	14	6	5	6	7	12	10	4	2	4	4	3	3
27	18	23	25	18	22	20	7	12	17	8	7	6	3	1	3	3	1	3
28	16	22	18	12	19	15	6	5	8	6	8	8	0	0	2	0	1	1
29	7	4	10	11	8	9	4	1	6	4	7	6	1	1	1	2	1	1
30	8	13	11	7	11	9	3	8	6	2	5	4	5	4	5	5	3	3

*denotes item with DIF

Proportions Correct	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.970	0.437	0.703	0.997	1.000	1.000
$\pi_1 = 0.30, \pi_2 = 0.70$	0.813	0.357	0.670	0.840	0.920	0.910
$\pi_1 = 0.50, \pi_2 = 0.50$	0.053	0.057	0.077	0.067	0.060	0.053

Proportions of False Positives	LCS	ESA	H	G^2	LR	MH
$\pi_1 = 0.17, \pi_2 = 0.83$	0.218	0.221	0.254	0.216	0.236	0.213
$\pi_1 = 0.30, \pi_2 = 0.70$	0.097	0.107	0.151	0.118	0.132	0.116
$\pi_1 = 0.50, \pi_2 = 0.50$	0.051	0.032	0.076	0.063	0.048	0.045

Table 6: Number of Times Each Item Was Identified as Exhibiting DIF using LC membership as the grouping variable ($\pi_1 = 0.50$, $\pi_2 = 0.50$; N=50)

	Latent Class Proportions							
	Δb =0.0 DIF Method		Δb = 0.3 DIF Method		Δb =1.0 DIF Method			
Item	LR	MH	LR	MH	LR	MH	$\bar{D}$-DIF	SD(D-DIF)
1	7	2	38*	37*	50*	50*	-1.9561	0.0727
2	3	5	43*	44*	50*	50*	-1.9502	0.0536
3	4	1	40*	36*	50*	50*	-1.9079	0.0416
4	2	4	4	1	50*	50*	-1.9459	0.0662
5	3	4	2	4	50*	50*	-1.9670	0.0810
6	2	2	2	1	50*	50*	-1.9420	0.0773
7	2	2	4	2	13	13	0.3891	0.1084
8	1	0	4	2	25	20	0.4309	0.0730
9	5	1	4	0	21	17	0.4208	0.0796
10	1	5	3	6	16	13	0.3463	0.0578
11	5	3	3	4	24	18	0.3914	0.0478
12	4	1	0	2	17	14	0.4344	0.0687
13	2	3	2	1	31	30	0.4827	0.0565
14	2	0	1	7	25	21	0.4018	0.0419
15	3	2	5	3	16	12	0.4089	0.0593
16	2	2	4	1	26	27	0.4197	0.0538
17	4	2	5	0	29	26	0.4488	0.0469
18	0	2	3	2	36	35	0.5615	0.0447
19	2	1	6	4	18	12	0.3424	0.0885
20	5	3	3	5	28	25	0.4430	0.0482
21	1	2	4	1	27	23	0.4736	0.0452
22	2	1	2	3	15	13	0.5616	0.1795
23	3	3	7	3	24	22	0.3867	0.0556
24	1	1	6	1	29	28	0.4937	0.0499
25	4	1	1	6	25	22	0.4156	0.0443
26	3	3	2	5	32	28	0.5081	0.0490
27	4	0	5	2	29	26	0.4715	0.0646
28	3	2	2	4	30	27	0.4676	0.0473
29	3	3	1	3	15	12	0.3496	0.0463
30	0	3	1	4	25	20	0.4063	0.0475

*denotes item with DIF

Proportions Correct ($\pi_1 = 0.50$, $\pi_2 = 0.50$)	LR	MH
Δb = 0.0	0.070	0.060
Δb = 0.3	0.430	0.410
Δb =1.0	1.000	1.000

Proportions of False Positives ($\pi_1 = 0.50$, $\pi_2 = 0.50$)	LR	MH
Δb = 0.0	0.055	0.043
Δb = 0.3	0.064	0.057
Δb =1.0	0.480	0.420

Figure 1: Conceptualization of the Relationship between Manifest Groups and Latent Classes

		Latent Classes A	B
Manifest Groups	Focal	100%	0
	Reference	0	100%

		Latent Classes A	B
Manifest Groups	Focal	80%	20%
	Reference	20%	80%

		Latent Classes A	B
Manifest Groups	Focal	50%	50%
	Reference	50%	50%

Figure 2: Design

		Latent Class Proportions					
		$\pi_1 = 0.17, \pi_2 = 0.83$		$\pi_1 = 0.30, \pi_2 = 0.70$		$\pi_1 = 0.50, \pi_2 = 0.50$	
		Δb		Δb		Δb	
		0.3	1.0	0.3	1.0	0.3	1.0
%	0%		x		x		x
DIF	10%	x	x	x	x	x	x
items	20%	x	x	x	x	x	x

Figure 3: Theoretical Crosstabulations for Manifest Groups and Latent Classes

Figure 3a:

		Latent Classes A	B	Total	Prop
Manifest Groups	Focal	500	0	500	0.17
	Reference	0	2500	2500	0.83
	Total	500	2500	3000	
	π_Vs	0.17	0.83		

Figure 3b:

		Latent Classes A′	B	Total	Prop
Manifest Groups	Focal	400	100	500	0.17
	Reference	500	2000	2500	0.83
	Total	900	2100	3000	
	π_Vs	0.30	0.70		

Figure 3c:

		Latent Classes A	B	Total	Prop
Manifest Groups	Focal	250	250	500	0.17
	Reference	1250	1250	2500	0.83
	Total	1500	1500	3000	
	π_Vs	0.50	0.50		

Acknowledgments

We wish to thank Cynthia Barton for allowing us to use the NELS cognitive test data.

U.S. Department of Education
Office of Educational Research and Improvement (OERI)
National Library of Education (NLE)
Educational Resources Information Center (ERIC)

TM031967

REPRODUCTION RELEASE

(Specific Document)

I. DOCUMENT IDENTIFICATION:

Title: A RECONCEPTUALIZATION OF DIFFERENTIAL ITEM FUNCTIONING	
Author(s): R.J. De Ayala, Seock-Ho Kim, Laura M. Stapleton, & C. Mitchell Dayton	
Corporate Source: Univ of Nebraska	Publication Date:

II. REPRODUCTION RELEASE:

In order to disseminate as widely as possible timely and significant materials of interest to the educational community, documents announced in the monthly abstract journal of the ERIC system, *Resources in Education* (RIE), are usually made available to users in microfiche, reproduced paper copy, and electronic media, and sold through the ERIC Document Reproduction Service (EDRS). Credit is given to the source of each document, and, if reproduction release is granted, one of the following notices is affixed to the document.

If permission is granted to reproduce and disseminate the identified document, please CHECK ONE of the following three options and sign at the bottom of the page.

The sample sticker shown below will be affixed to all Level 1 documents	The sample sticker shown below will be affixed to all Level 2A documents	The sample sticker shown below will be affixed to all Level 2B documents
PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 1	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE, AND IN ELECTRONIC MEDIA FOR ERIC COLLECTION SUBSCRIBERS ONLY, HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2A	PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY Sample TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) 2B
Level 1 ↑ [X]	Level 2A ↑ [ ]	Level 2B ↑ [ ]
Check here for Level 1 release, permitting reproduction and dissemination in microfiche or other ERIC archival media (e.g., electronic) *and* paper copy.	Check here for Level 2A release, permitting reproduction and dissemination in microfiche and in electronic media for ERIC archival collection subscribers only	Check here for Level 2B release, permitting reproduction and dissemination in microfiche only

Documents will be processed as indicated provided reproduction quality permits.
If permission to reproduce is granted, but no box is checked, documents will be processed at Level 1.

I hereby grant to the Educational Resources Information Center (ERIC) nonexclusive permission to reproduce and disseminate this document as indicated above. Reproduction from the ERIC microfiche or electronic media by persons other than ERIC employees and its system contractors requires permission from the copyright holder. Exception is made for non-profit reproduction by libraries and other service agencies to satisfy information needs of educators in response to discrete inquiries.

Sign here, → please	Signature: Ralph De Ayala	Printed Name/Position/Title: RALPH DE AYALA / ASSOC PROF	
	Organization/Address: Univ of Nebraska, Ed Psyc, Lincoln, NE 68588	Telephone: 402 472 2736	FAX:
		E-Mail Address: rdeayala2@unl.edu	Date:

(over)

III. DOCUMENT AVAILABILITY INFORMATION (FROM NON-ERIC SOURCE):

If permission to reproduce is not granted to ERIC, or, if you wish ERIC to cite the availability of the document from another source, please provide the following information regarding the availability of the document. (ERIC will not announce a document unless it is publicly available, and a dependable source can be specified. Contributors should also be aware that ERIC selection criteria are significantly more stringent for documents that cannot be made available through EDRS.)

Publisher/Distributor:
Address:
Price:

IV. REFERRAL OF ERIC TO COPYRIGHT/REPRODUCTION RIGHTS HOLDER:

If the right to grant this reproduction release is held by someone other than the addressee, please provide the appropriate name and address:

Name:
Address:

V. WHERE TO SEND THIS FORM:

Send this form to the following ERIC Clearinghouse:

University of Maryland
ERIC Clearinghouse on Assessment and Evaluation
1129 Shriver Laboratory
College Park, MD 20742
Attn: Acquisitions

However, if solicited by the ERIC Facility, or if making an unsolicited contribution to ERIC, return this form (and the document being contributed) to: